The geographic harmonisation of Scotland's small area census data, 1981 to 2011

The geographic harmonisation of Scotland's small area census data, 1981 to 2011

Daniel J Exeter,[1] Zhiqiang Feng,[2] Jinfeng Zhao,[1] Alana Cavadino,[1] Paul Norman[3]

Affiliations

[1] School of Population Health, The University of Auckland, Auckland, New Zealand

[2] School of Geosciences, The University of Edinburgh, Edinburgh, UK

[3] School of Geography, the University of Leeds, Leeds UK.

Abstract

Previous research in Scotland used a merging approach to combine census boundary data from 1981, 1991 and 2001 to create Consistent Areas Through Time (CATTs) for the analysis of health and social data for small areas. In this paper, we adopt the same methodology to integrate the 2011 Scottish Output Areas to the CATTs. First, we merged the 2001 and 2011 Output Areas to create SUPER OAs, which were then merged with SUPER EDs, which represented a consistent small area geography for 1981 and 1991. This resulted in 8,548 CATTs providing a consistent geography for the 1981, 1991, 2001 and 2011 Censuses in Scotland. We demonstrate the utility of the CATTs by exploring the correlations between deprivation, the proportion of the population who were permanently sick, and those with degree qualifications, across the 4 censuses. We have provided a resource that enables users to deepen their understanding of small area social changes in Scotland between the 1981 and 2011 Censuses.

Keywords: geographic harmonisation; CATTs; Scotland; Zone design; consistent geographies, permanently sick, degree qualifications

Introduction

The 2011 Census data for Scotland provides an opportunity for social scientists to investigate Scotland's contemporary socio-demographic structure and how its demography has changed since previous census time points. The 2011 Census night population of 5,295,403 was the largest population ever for Scotland, representing a growth of 4.6% since the 2001 Census and a 7.2% increase from the 4,939,503 people in Scotland for the 1981 Census. While researchers can describe national trends over time easily, changes to the configuration of small area boundary files for every census, in response to population growth or decline or perhaps due to subtle changes in the digitisation accuracy, restricts the ability to analyse changes over time at a more local scale (Norman et al. 2007).

Consequently, this has led to a number of approaches to harmonise two or more geographical datasets, to facilitate the analysis of social change in small areas. Common approaches include areal-weighting (Flowerdew and Green 1992), dasymetric mapping (e.g. (Syphard et al. 2009) and the conversion of data from irregular polygons into regular grids (Martin 1996; Lloyd et al. 2017). Norman et al (2003) outlined four alternative approaches to managing incongruent spatial units over time. These include: 'freezing' the geographical base, by selecting the spatial units at one point in time; 'transforming' historical data into contemporary zones; using individual or household level data and aggregating to meaningful zones; and developing designer zones common to all years.

Using postcode-level histories to monitor small area census boundary changes in Scotland between the 1981 and 1991 Censuses and then between the 1991 and 2001 Censuses, Exeter et al. (2005) used a merging approach to create Consistent Areas Through Time (CATTs). This is broadly similar to Norman's suggestion of developing designer zones, although the CATTs did not require any estimation in the geographic conversion process.

In this paper, we adopt Exeter et al. (2005)'s merging approach to update the CATTs to include the 2011 Census geography. The Methods section comprises the majority of this paper but in the Results, to demonstrate the utility of having consistently defined geographic units, we present changes in premature mortality for years around each of the 1981, 1991, 2001 and 2011 Censuses.

Methods

Small area census spatial data files

There were 17,767 Enumeration Districts (EDs) used for the collection and output of the 1981 Scottish Census data. Of these, 56 were 'shipping' EDs (one for each of the 56 Districts), leaving 17,711 residential EDs. The 1981 ED boundaries were not digitised, but were assigned population-weighted grid references. In 1991 there were 38,254 Output Areas (OAs) represented as population-weighted grid references. In addition, there were 38,098 digitised boundaries, (the remaining 156 OAs were large communal establishments and only represented as points). In 2001 there were 42,604 OAs represented both as population-weighted grid references and boundaries. In 2011, the number of OAs had increased to 46,351 that were available as population-weighted grid references and a polygon boundary file. A simple assessment of the number of areas at each census suggests that analysing change over time for small areas would be problematic, as it is difficult to establish which areas were split, merged or had their boundaries 'nudged' between censuses.

Residential address grid references

We obtained the 2011 Address Register file from National Records for Scotland (NRS), which contained grid references for the 2,429,647 residential addresses in Scotland in April 2011.

Census data

Population data were obtained in 5 year age bands for the original small area geographies from CASWEB (1981-2001) and InFuse (2011). We also obtained numerators and denominators from each census to calculate the proportion of the adult population who were 'permanently sick', and the proportion of the population who had a university degree. For each variable separately, we combined or 'stacked' the data into a single file and calculated k-means clusters representing 5 'permanently sick' groups and 5 'degree' groups, representing patterns of change for the consistent areas across the 4 census years. The 5 'permanently sick' clusters represented the areas that were: Always advantaged; Largely worsening; Greatly improving; Largely improving; and Somewhat worsening. The 5 'degree' clusters classified areas that were: Mixed picture; Fair and increasing; Highly educated; Low but improving; and Substantially improving.

We also data to calculate the Carstairs Index for each census period, which we categorised into population-weighted quintiles for each census year. We also 'stacked' the Carstairs data for each census upon each other and created k-means clusters of deprivation, which distinguished CATTs into 5 groups representing areas which were persistently deprived, deprived but improving, persistently average deprivation, moderately advantaged but worsening, and persistently not deprived.

Statistical analyses

We used Spearman rank correlations to explore the associations between the K-means clusters representing trajectories of the permanently sick, degree, and deprived populations over time with the data from each census separately. These statistical analyses were conducted in Stata version 15.

Creating Consistent Areas Through Time linking 1981, 1991, and 2001 small area output geographies

Exeter et al. (2005) provide a detailed account of constructing the CATTs for 1981 to 2001. Essentially, the process was to create a digital boundary file of the 1981 Enumeration Districts, which were an aggregation of the 1991 Output Areas. However, dissolving the 1991 OAs only created 16,096 'pseudo EDs' so the Central Postcode Directory was used to correctly allocate the remaining 1,615 EDs that had been absorbed into the 1991 OA geography. Overlapping zones were merged which resulted in the creation of 15,739 polygons ('SUPER EDs') linking the 1981 EDs with the 1991 OAs.

Geographic Conversion Tables (GCTs) assigned the 1981 EDs and 1991 OAs to the SUPER EDs. Unlike existing GCTs(Simpson 2002; Norman et al. 2003), an estimated apportionment 'weight' is not needed since the 1981 EDs and 1991 OAs nest wholly within the SUPER EDs. The 2001 OA polygons were merged with the SUPER ED polygons and 2001 AddressPoint™ data were used to distinguish polygons of genuine boundary changes from sliver polygons resulting from digitisation differences of the 1991 and 2001 OA boundaries. This resulted in 10,058 polygons integrating the 1981, 1991 and 2001 small area boundary files. GCTs linked the 1981 EDs, 1991 OAs and 2001 OAs to the CATTs which were made available from the Census Data Users website at MIMAS and UKBORDERS from EDINA.

Incorporating the 2011 Census to the existing Consistent Areas Through Time

1. *Merging 2001 and 2011 Output Areas*

Enhancing the existing CATTs and merging the 2011 OAs with the previous polygons would not suffice since it would not be possible to identify those 2001 OAs that were split or merged in the creation of the 2011 OAs by the NRS.Leveraging the topological information from ArcInfo coverages, we used a two-stage process in which ArcGIS v10.4 was used to first merge the 42,604 OAs for 2001 with the 46,351 OAs for 2011 resulting in a file 182,479 unique polygons. We obtained a count of AddressPoints™ in each polygon and consistent with Exeter et al. (2005) those polygons with ≤2 AddressPoints™ were treated as sliver polygons and removed. When eliminating the sliver polygons, we ensured that the boundary of the 2011 OA was retained, so that our final output would conform to boundaries of the contemporary OAs and other higher administrative geographies. This process

resulted in the 36,921 unique "SUPER OAs". Note that our Scottish SUPER OAs are not official intermediate geographies.

2. *Merging SUPER OAs and SUPER EDs*

We repeated the merge process and combined the 36,921 SUPER OAs with the 15,739 SUPER EDs and used the 2011 AddressPoints™ to identify sliver polygons using the ≤2 AddressPoints™ rule. This process resulted in 8,365 Consistent Areas Through Time for 1981 to 2011. In 2011, the mean CATT population was 633, but the maximum was 95,852. After relaxing the definition of a sliver polygon to comprise polygons ≤5 AddressPoints™, resulting in 8,557 CATTs. Quality checking found some OAs along the coastline or bodies of inland water overlapped more than one CATT, which required manual correction, resulting in 8,548 CATTs.

CATTs were labelled according to the lowest common SUPER ED code, with a "C2" or "C5" prefix to specify the number of AddressPoints™ used to identify sliver polygons. The 2011CATT shapefile and the GCTs linking 1981 EDs, and OAs from 1991, 2001 and 2011 to CATTs were prepared and are freely available from the National Records for Scotland, Scottish Longitudinal Studies and Information and Statistics Division of the NHS websites.

For the resident population, while there was relative consistency in the decreasing mean and standard deviations for the official geographies over time (Table 1), with a notable decrease between 1981 and 1991 when the minimum population thresholds increased to 50 per OA. By contrast, there were stepwise increases in the mean, maximum and standard deviations by CATT population (Table 1).

Population	1981		1991		2001		2011	
	1981 EDs	2011 CATTs	1991 OAs	2011 CATTs	2001 OAs	2011 CATTs	2011 OAs	2011 CATTs
N	17711	8548	38253	8548	42604	8548	46351	8548
Min	0	40	46	50	50	51	50	51
Max	1332	37722	773	39409	2357	43936	2081	55807
Mean	283.75	587.92	130.69	584.78	118.82	592.19	114.25	619.49
Std. Dev.	146.33	999.71	46.6	1061.91	44.96	1218.48	45.07	1428.35
Total	5 025 516*		4 999 303*		5 062 011		5 295 403	

Table 1: Summary statistics for the Total Usual Resident population between 1981 and 2011, by Official small area outputs and CATTs *excludes Shipping zones

Results

Figure 1 shows the distribution of the K-Means clusters representing trajectories of (a) the permanently sick population, (b) those with Degree qualifications and (c) deprivation, between 1981 and 2011. The overwhelming majority of rural CATTs in Scotland are always advantaged, with relatively few people in each census stating they were permanently sick. Surprisingly, rural areas – particularly the remote islands, in the central belt and in the rural South are also areas that were somewhat worsening over time. The number of CATTs that were greatly improving is few (N23) while there are few areas typically on the outskirts of the main centres largely improving.

In terms of trajectories for the population with a university degree, vocation or professional certificate, the distrubtion of this is more disparate than those seen for the permanently sick. While the highly educated group tends to cluster in main centres or University towns, there is an overall pattern of improving levels of education in most parts of the country, particularly in the central belt. The rural and remote areas in the Highlands and Islands are more likely to experience Fair and improving conditions, indispersed with pockets of 'mixed picture' CATTs, representing flucations in the population with degree level education from one census to the next.

The K-means clusters of deprivation demonstrates that those areas that are persistently deprived are synonymous with CATTs that were in the North West of Scotland, with areas of high deprivation but improving located in and around Glasgow. The CATTs that were experiencing persistently average deprivation circumstances were scattered throughout the rural and urban parts of Scotland, particularly in Dundee, Edinburgh and Aberdeen, with a scattering in the South-West and North Eastern areas. Many of the "persistently advantaged" (and moderately ad but worsening) CATTs are located in Edinburgh, with a considerably fewer in Glasgow, which exhibits a more of a mixed distribution of deprivation in general.

Table 2 shows the correlations between the continuous measures of permanently sick, population with degrees, and deprivation from each census period between 1981 and 2011. The inter-census correlations for the %permanently sick were moderate-strong overall, however the association waned over time. For example the %permanently sick in 1981 was strongest with the corresponding variable in 1991(0.60), decreasing to 0.52 for %permanently sick in 2011. Similarly, the %permanently sick in 1991 was associated most with the 2001 %permanently sick (0.81) and the strongest correlation overall for %permanently sick was between 2001 and 2011 (0.83).

Within the %degree results, the strongest correlations were also between 2001 and 2011, (0.93) although the 1981 %degree correlations followed an 'n' distribution, peaking at 0.68 with the 2001

%degree. The correlations between the continuous Carstairs score over time scores were the strongest and most consistent overall, ranging from 0.80 (between 1981 and 2001) and peaking at 0.86 between 2001 and 2011.

There were statistically significant associations between all pairs of variables ($P<0.001$ for all comparisons), with these correlations ranging from moderate to strong. On average, CATTs areas with higher levels of education had lower proportions of permanently sick reported, , with this negative correlation strengthening over time from -0.42 in 1981 to a peak of -0.75 in 2001, with a marginal reduction to -0.72 in 2011. Higher levels of education were also correlated with lower deprivation scores, although the strength of this negative association followed a more erratic pattern across these four census years. Increased percentages of permanently sick were associated with higher deprivation scores; this correlation was moderate in 1981, but strengthened over time, also peaking in 2001 before weakening slightly in 2011.

Table 2. Spearman's correlation coefficient for associations between CATTs level continuous measures of percent permanently sick, population with university degrees, and deprivation index from each census period between 1981 and 2011.

	% permanently sick				% with Degree				Deprivation index			
	1981	1991	2001	2011	1981	1991	2001	2011	1981	1991	2001	2011
% permanently sick												
1981	1											
1991	0.60	1										
2001	0.57	0.81	1									
2011	0.52	0.73	0.83	1								
% with Degree												
1981	-0.40	-0.54	-0.58	-0.56	1							
1991	-0.36	-0.52	-0.58	-0.57	0.59	1						
2001	-0.45	-0.68	-0.75	-0.73	0.68	0.72	1					
2011	-0.42	-0.66	-0.72	-0.72	0.64	0.69	0.93	1				
Deprivation index												
1981	0.52	0.71	0.73	0.71	-0.66	-0.57	-0.73	-0.69	1			
1991	0.53	0.71	0.74	0.73	-0.60	-0.56	-0.70	-0.64	0.84	1		
2001	0.50	0.70	0.76	0.77	-0.65	-0.62	-0.81	-0.77	0.84	0.85	1	
2011	0.50	0.68	0.73	0.74	-0.57	-0.53	-0.70	-0.66	0.80	0.84	0.86	1

Note. All correlations reported are statistically significant with P<0.001

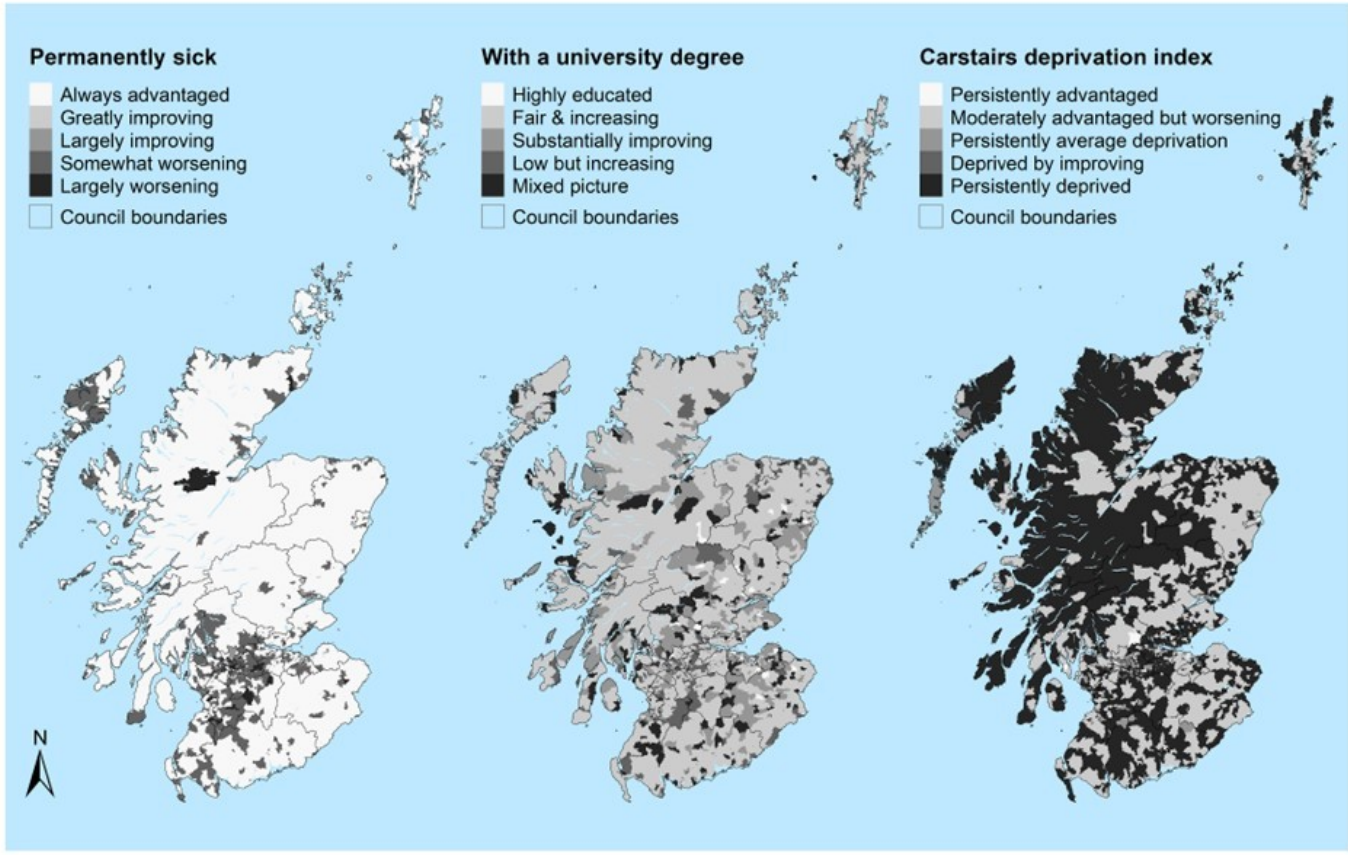

Figure 1: K-Means clusters representing trajectories of (a) the permanently sick population, (b) those with degree qualifications and (c) the Carstairs deprivation index. Note that each set of clusters uses census data from 1981, 1991, 2001 and 2011 combined.

Discussion

This paper describes the extension of Consistent Areas Through Time to facilitate analyses of health and social data in Scotland between 1981 and 2011 for small areas. The previous CATTs linking 1981, 1991 and 2001 Census Enumeration Districts and Output Areas comprised 10,058 areas and had a mean population of approximately 500. Introducing the 2011 Census OAs reduced the number of zones to 8,548, while the mean population remained broadly similar increasing to 619 in 2011. A strength of the methodology used is that population estimation is not required when aggregating data from ED or OAs to the CATT. However, a limitation is that a small proportion (0.08%) of CATTs have populations in excess of 20,000.

Our research contributes to the literature concerning the need to have consistent geographical units over time. Accepted best practice for the comparison of population group with differing population distributions is to age (or age and sex) standardise the data (Rothman 2008). Yet, many studies assessing changes in health outcomes or social conditions over time typically use period-specific measures of those health or social outcomes (e.g. Phillimore et al. 1994; Tobias et al. 2008; Allik et al. 2016; McCartney et al. 2017). This means, for example, a study comparing the association between area deprivation and mortality between 1981 and 2011 would use an area deprivation measure derived from the 1981 census to calculate mortality rates for that period and use the 2011 census-based deprivation index for the more recent period. Such results demonstrate how the social gradient has changed over time, but the ability to study areas whose socio-economic conditions have improved, worsened or remained broadly consistent over time is impossible (Norman et al. 2011). Previously, we found that although premature mortality had reduced in all period-specific deprivation quintiles in 1981 and 2001, when we used time comparable deprivation in the CATTs approach, mortality in the persistently deprived areas increased by nearly 10% over the 20 year period (Exeter et al. 2011).

In this study, we used data from the 1981, 1991, 2001 and 2011 census regarding the population who were permanently sick, or who had a university degree-level education as well as the Cartairs index of deprivation, aggregated to the CATT level in order to demonstrate the benefits of having consistent geographical areas over time. The K-means clusters we created here use data from all four time periods combined to produce a classification schema that could not be achieved had we used the data for each time period separately.

The use of K-means clusters is not uncommon in population health research. For example, Green et

al (2014) created a neighbourhood classification scheme based on mortality from 63 outcomes
11 between 2006 and 2009 in the UK. Their results demonstrated that the clustering of areas into a typologies such as 'Best health and most desirable', 'Poor Health Experiences' or 'Mixed Experiences' can inform policy development, resource allocation and targeting of services. Similarly, Bellis et al (2012) used the K-means method to group 30 public health metrics commonly used in the UK into 5 classes and to describe how areas in England were differently affected by various factors across the life course.

A study such as this is not without its limitations. First, while the geographic areas are consistent over time, we do not know about the migration patterns of the population between each census. Nevertheless, such studies are possible by using the CATTs within the Scottish Longitudinal Study's data environment. This would enable us to further compare the outcomes we used in this study for individuals who themselves did not move over time, or experienced individual-level social mobility. Second, we acknowledge that the census questions from which our permanently sick, degree-level education and Carstairs index of deprivation are derived have not been consistent over time. In some cases, the approach to asking the question in the Census has changed. For example, in earlier censuses, the question regarding higher education asked respondents to write the details of their qualifications down, but this changed in 2001 to ask people to tick whether they had either/or a degree, a professional certificate and vocational certificate. There has also been a sea-change in the way occupational social classes have been classified by the National Records for Scotland, consistent with the Office for National Statistics (Rose et al. 2005).

Researchers, and government and non-government organisations have used the original CATTs to explore geographic variations in health society between 1981 and 2001 (Platt et al. 2007; Popham et al. 2010; Walsh 2014). To have data on this consistent geographic basis over time, raw data for the original ED or OA geography are aggregated using lookup tables to the CATTs units. Through these lookup tables and associated GIS boundary files we have provided a resource that enables researchers to deepen their understanding of small area social changes in Scotland between the 1981 and 2011 Censuses. The CATTs are freely available for download from (URLS TO BE ADDED).

References

Exeter DJ, Boyle P, Feng Z, Flowerdew R, Schierloh N. 2005. The creation of 'consistent areas through time' (CATTs) in Scotland, 1981-2001. Population trends.(119):28-36.
Flowerdew R, Green M. 1992. DEVELOPMENTS IN AREAL INTERPOLATION METHODS AND GIS. Annals of Regional Science. 26(1):67-78.

Lloyd CT, Sorichetta A, Tatem AJ. 2017. High resolution global gridded data for use in population studies. Scientific Data. 4.
Martin D. 1996. An assessment of surface and zonal models of population. International Journal of Geographical Information Systems. 10(8):973-989.
Norman P, Purdam K, Tajar A, Simpson L. 2007. Representation and local democracy: Geographical variations in elector to councillor ratios. Political Geography. 26(1):57-77.
Norman P, Rees P, Boyle P. 2003. Achieving data compatibility over space and time: Creating consistent geographical zones. International Journal of Population Geography. 9(5):365-386.
Platt S, Boyle PJ, Crombie I, Feng Z, Exeter DJ. 2007. The epidemiology of suicide in Scotland 1989-2004: an examination of temporal trends and risk factors at national and local levels. Edinburgh: Scottish Executive Social Research.
Popham F, Boyle P, O'Reilly D, Leyland AH. 2010. Exploring the impact of selective migration on the deprivation-mortality gap within Greater Glasgow. Glasgow: Glasgow Centre for Population Health.
Simpson L. 2002. Geography conversion tables: a framework for conversion of data between geographical units. . International Journal of Population Geography. 8(1):69-82.
Syphard AD, Stewart SI, McKeefry J, Hammer RB, Fried JS, Holcomb S, Radeloff VC. 2009. Assessing housing growth when census boundaries change. International Journal of Geographical Information Science. 23(7):859-876.
Walsh D. 2014. An analysis of the extent to which socio-economic deprivation explains higher mortality in Glasgow in comparison with other post-industrial UK cities, and an investigation of other possible explanations. Glasgow: University of Glasgow.